基于 EK-medoids 聚类和邻域距离的特征选择方法 *

孙印杰 [a]，张新乐 [a]，孙　林 [a, b †]

(河南师范大学 a. 计算机与信息工程学院; b. 河南省高校计算智能与数据挖掘工程技术研究中心, 河南 新乡 453007)

摘　要：针对传统聚类算法中只注重数据间的距离关系，而忽视数据全局性分布结构的问题，提出一种基于 EK-medoids 聚类和邻域距离的特征选择方法。首先，用稀疏重构的方法计算数据样本之间的有效距离，构建基于有效距离的相似性矩阵；然后，将相似性矩阵应用到 K-medoids 聚类算法中，获取新的聚类中心，进而提出 EK-medoids 聚类算法，可有效对原始数据集进行聚类；最后，根据划分结果所构成簇的邻域距离给出确定数据集中的属性重要度定义，应用启发式搜索方法设计一种 EK-medoids 聚类和邻域距离的特征选择算法，降低了聚类算法的时间复杂度。实验结果表明，该算法不仅有效地提高了聚类结果的精度，而且也可选择出分类精度较高的特征子集。

关键词：特征选择；有效距离；K-medoids 聚类；邻域距离

中图分类号：TP301.6　　doi: 10.3969/j.issn.1001-3695.2018.02.0093

Feature selection method based on EK-medoids cluster and neighborhood distance

Sun Yinjie[a], Zhang Xinle[a], Sun Lin[a, b†]

(*a. College of Computer & Information Engineering, b. Engineering Technology Research Center for Computing Intelligence & Data Mining of Henan Province Henan Normal University, Xinxiang Henan* 453007, *China*)

Abstract: Since the traditional clustering algorithms only pay attention to the distance relationship among data, and ignore the problem of global distribution data structure, this paper proposed a feature selection method based on EK-medoids cluster and neighborhood distance. First of all, it calculated the effective distances between data samples by using the sparse reconstruction method, and constructed an effective distance-based similarity matrix. Then it matrixed the similarity introduced in the K-medoids clustering algorithm, and obtained these new cluster centers. This paper developed an EK-medoids clustering algorithm which can effectively cluster these original data sets. Finally, it investigated a neighborhood distance in neighborhood rough set, and according to the classification results of clusters, it defined an attribute importance based on the neighborhood distance, and designed an EK-medoids cluster and neighborhood distance-based feature selection algorithm on the basis of heuristic searching method, which can further reduce the time complexity of cluster algorithms. The experimental results show that our proposed algorithm not only effectively can improve the accuracy of the clustering results but also select the feature subset with high classification accuracy.

Key words: feature selection; effective distance; K-medoids cluster; neighborhood distance

0　引言

微阵列技术是一种研究基因表达的技术，通过分析基因表达谱数据中成千上万的基因数据而得到有价值的信息，目前已经广泛的应用到医学等各个领域。与疾病有关的基因表达谱数据分类已经成为生物医学研究领域的一个重要研究方向[1]。近年来，随着科学技术的不断发展，基因表达谱数据量急速增长，并表现出规模庞大、内容复杂的特性，一方面特征空间的维数不断增加而降低了学习算法的效率，另一方面大量冗余数据的出现干扰了实验的结果[2]。为了降低这些不利因素造成的影响，诸多学者提出了许多的特征（基因）选择算法[3~8]。

聚类分析长期以来在各个领域扮演着重要的作用，包括金融、医疗、图像、和生物信息学等方面[9-15]。聚类方法通常并不需要使用训练数据进行学习，因此该类方法属于无监督学习的

收稿日期：2018-02-27；**修回日期：**2018-04-09　**基金项目：**国家自然科学基金资助项目（61772176，11702087）；中国博士后科学基金资助项目（2016M602247）；河南省科技创新人才项目（184100510003）；河南省科技攻关项目（182102210362，162102210261，182102210078）；河南省高校青年骨干教师培养计划资助项目（2017GGJS041）；河南省自然科学基金资助项目（182300410130）；河南省高等学校重点科研计划资助项目（14A520069，17A520038，16A520015）；新乡市科技攻关计划资助项目（CXGG17002）；河南师范大学博士科研启动费支持课题（qd15132，qd15129，qd15131）；河南师范大学青年科学基金资助项目（2015QK23，2015QK24）

作者简介：孙印杰（1963-），男，河南南阳人，教授，硕士，主要研究方向为数据挖掘、多媒体技术等；张新乐（1993-），男，河南南阳人，硕士研究生，主要研究方向为粒计算、数据挖掘等；孙林（1979-），男（通信作者），河南南阳人，副教授，博士，主要研究方向为粒计算、大数据挖掘、生物信息学等（sunlin@htu.edu.cn）.

范畴。近年来，随着社会信息化水平越来越高，需要处理的数据也越来越多。数据的高维度使数据具有稀疏、不可聚集等特性，使得大量的聚类算法在处理高维空间数据时并不尽人意[16-20]。特征选择对剔除冗余、无关的属性，减少后续算法的时间复杂度，提高分类精度，精简算法有着非常重要的作用[21-22]。因此，优化聚类算法，并将聚类分析与特征选择算法相结合，进而设计高效的特征选择模型及算法是非常有必要的。

对于上面的问题，许多学者已经做出了大量的研究来减少高维度对数据提取的影响，常用的方法有基于粗糙集的特征选择算法、基于遗传算法的特征选择算法、基于模式相似性判断和信息增益的特征选择算法[3-6, 22-26]。例如，胡清华等研究了邻域粗糙集理论，并将其应用于特征选择算法，处理连续性数据[24]。段洁等人针对多标记分类任务，重新定义了邻域粗糙集下近似和依赖度的计算方法[25]。孙林等结合局部线性嵌入算法和邻域粗糙集模型，提出了一种基因表达谱数据的基因选择方法[26]。

传统的距离函数包括马氏距离、欧氏距离、切氏距离、明视距离、归一化距离和绝对值距离等，这些距离函数只关注数据之间的地理距离，虽然在计算时简单，但忽略了其它有价值的信息（如拓扑几何关系等）[23]。为此，本文引入有效距离和邻域距离，提出了一种基于 EK-medoids 聚类和邻域距离的特征选择方法。该算法首先将有效距离应用于聚类算法中，对原始数据聚类，然后与邻域系统中定义的邻域距离结合，根据聚类所划分簇的邻域距离，计算属性重要度，并在此基础上利用启发式搜索方法设计特征选择算法。

1　邻域粗糙集

Pawlak[27]于 1991 年提出了粗糙集理论，它可以有效地处理不精确或模糊的概念。粗糙集理论及其应用发展迅速，已成为一种处理不确定数据并进行特征选择、规则提取和知识发现的有效工具[28,29]。目前，特征选择方法使用较多的主要有基于过滤算法的方法和基于封装算法的方法两类[7]。由于过滤法的评价准则为数据之间的特性，因而所选特征之间相关性比较强。封装法的评价准则通过特定分类器来实现，在特征选择过程中需要多次调用分类算法，进而导致了算法的时间复杂度较高[24]。传统的粗糙集理论虽然选用了等价类形式化地表示了知识分类，然而这些等价类显然是通过划分获得的，对于连续型数据的等价类，其离散化的过程必定会导致某类关键信息的丢失[25]。邻域粗糙集能够有效地弥补经典粗糙集理论的上述缺点。下面基于邻域关系简要介绍邻域粗糙集的一些相关概念[4, 24~26]。

给定一个邻域信息系统 $IS = (U, A, V, f, \delta)$，该五元组中的 U 为非空有限集，称为论域，A 为特征集，特征 a 的值域为 V，$V = \bigcup_{a\in A} V_a$，其中 V_a 表示特征 a 的值域，该邻域信息系统的信息函数为 $f: U \times A \rightarrow V$，即对任意 $x\in U$ 且 $a\in A$，有 $f(x, a)\in V_a$，邻域信息系统的阈值为 $\delta\in[0, 1]$。

给定一个五元组 $IS = (U, A, V, f, \delta)$中任意 $x, y\in U$，$B\subseteq A$，$B = \{a_1, a_2, \ldots, a_n\}$，$B$ 上的距离函数为 $D_B(x, y)$，其表达式为

$$D_B(x,y)=(\sum_{i=1}^{n}(|f(x,a_i)-f(y,a_i)|)^p)^{\frac{1}{p}} 。\quad (1)$$

当 $p=1$ 时，式(1)为曼哈顿距离；当 $p=2$ 时，式(1)为欧氏距离。

给定一个邻域信息系统 $IS = (U, A, V, f, \delta)$，$\forall x\in U$，$B\subseteq A$，$n_B^{\delta}(x)$ 被定义为 x 在 B 上的 δ 邻域，其表达式为：

$$n_B^{\delta}(x)=\{y\mid x,y\in U,D_B(x,y)\le\delta\} 。\quad (2)$$

由上面距离函数的定义可知，邻域 $n_B^{\delta}(x)$ 必须满足以下条件：

a) $n_B^{\delta}(x)\neq\varnothing$；

b) $x\in n_B^{\delta}(x)$；

c) $y\in n_B^{\delta}(x)\Leftrightarrow x\in n_B^{\delta}(y)$；

d) $\bigcup_{x\in U} n_B^{\delta}(x)=U$ 。

给定一个邻域信息系统 $IS = (U, A, V, f, \delta)$，任意特征子集 $B\subseteq A$ 决定了一个邻域阈值 δ 上的邻域关系为 $NR_\delta(B)$，其表达式为

$$NR_\delta(B)=\{(x,y)\in U\times U\mid D_B(x,y)\le\delta\} 。\quad (3)$$

根据式(3)可以得到 U 的邻域划分为 $U/NR_\delta(B)$，进而可以定义 U 上的一簇邻域知识。以此类推，邻域信息系统 IS 中的每个邻域划分称为一个邻域类或者邻域知识，可以得出上述 $n_B^{\delta}(x)$ 就是一个邻域类。

2　基于 EK-medoids 聚类和邻域距离的特征选择

2.1　基于有效距离的相似性矩阵

Brockmann 等人在寻找严重危害人类健康的疾病传播因素和途径时，提出了一种基于有效距离的度量函数[30-31]。通过大量数据实验表明，该函数能够有效模拟出 SARS 病毒和 H1N1 病毒在全球传播的情况，与传统的距离度量相比，有效距离度量能够利用数据样本之间的全局性结构信息，而降低数据的样本分布、地理距离等不良因素的影响[30]。高效的数据要能够有效地表示出来，稀疏表示从数据自身学习到几个典型（原子）模式的一个组合（通常为线性的），可以有效地表达出高效数据的全局特性。

假设有一批样本数量为 n，样本维数为 d 的样本集 $X=[x_1, x_2,\ldots,\ldots x_n]^{\mathrm{T}}\in R^{n\times d}$，要分成 k 个类，w 是稀疏表示时得到的

权重系数，P 表示归一化的权重系数，ED 表示有效距离，依据文献[31]分别给出权重系数矩阵、归一化的权重系数矩阵和有效距离矩阵。

a) 在数据样本之间，从稀疏表示的过程中所占的权重系数构建有向图，进而得到权重系数 w_i，其表达式为：

$$\min_{w_i} \| x_i - Bw_i \|_2^2 + \lambda \| w_i \|_1 , \tag{4}$$

其中，$w_i \geq 0$，x_i 代表第 i 个训练样本；$B = [\ ]$是一个 $d \times n$ 的矩阵，它的每一行是一个属性，每一列是一个样本，B 中所包含的是所有训练样本中除了 x_i 之外的全部样本，$B = [x_1, x_2, \ldots x_{i-1}, x_{i-2}, \ldots x_n]^{\mathrm{T}}$ 表示样本 x_i 从 X 中移除。w_{ij} 表示样本 x_i 在稀疏表示样本 x_j 时前面的系数；λ 是稀疏表示过程中的正则化参数 $\lambda \in (0, 1]$，λ 越大则矩阵越稀疏。

根据式(4)计算求得权重系数矩阵 $W = [w_1, w_2, \ldots, w_n]^{\mathrm{T}}$，该矩阵是一个 $n \times n$ 的权重矩阵。

b) 数据样本间归一化后的权重系数为 p_{ij}，其表达式为

$$p_{ij} = \frac{w_{ij}}{\sum_{i=1}^{n} w_{ij}} 。 \tag{5}$$

根据式(5)得到归一化的权重系数矩阵 P。如果 p_{ij} 越大，则说明 x_i 在稀疏重建 x_j 时，所占的权重越大，也就是表示 x_i 在 x_j 的所有近邻中位置更靠前，x_i 与 x_j 之间的相似度越大，有效距离就越小。

c)计算样本的有效距离，进而得到有效距离矩阵 ED，其表达式为

$$ED_{ij} = 1 - lnP_{ij} 。 \tag{6}$$

由于存在 $0 \leq p_{ij} \leq 1$，则 $lnP_{ij} \leq 0$，于是可得 $ED_{ij} \geq 1$。

定义 1 利用稀疏表示的方法计算原始数据之间的有效距离矩阵 $ED \in R^{n \times n}$，基于有效距离矩阵构造相似性矩阵 A，$A = [\ ]$是一个 $n \times n$ 的矩阵且 $A \in R^{n \times n}$，则有 a) 如果 $i \neq j$，则 $A_{ij} = exp(-ED_{ij} * ED_{ji})$；b)如果 $i = j$，则 $A_{ij} = 0$。

2.2 基于相似性矩阵的 K-medoids 聚类

传统的 K-medoids 聚类算法通常使用欧氏距离[32]，本文结合基于有效距离的相似性矩阵，改进 K-medoids 聚类算法，提出基于相似性矩阵的 K-medoids（EK-medoids）聚类方法。

传统的 K-medoids 聚类算法中，计算第 i 个聚类簇的新聚类中心的方法为

$$\max_{r_{ij}} \sum_{q=1}^{m} f(x_{r_{ij}}, x_{r_{iq}}) = \max_{r_{ij}} \sum_{q=1}^{m} \exp(-\| x_{r_{ij}} - x_{r_{iq}} \|_2^2) \tag{7}$$

其中:m 表示该聚类簇所包含的数据样本的个数；r_{jq} 表示属于该聚类簇的第 q 个数据样本在原始数据样本中的编号。

式(7)中的目标是寻找到最优对($x_{r_{ij}}, x_{r_{iq}}$)，进而使目标函数 $f(x_{r_{ij}}, x_{r_{iq}})$取得最小值。

定义 2 在 EK-medoids 算法中，计算聚类中心的公式为：

$$\max_{r_{ij}} \sum_{q=1}^{m} f(x_{r_{ij}}, x_{r_{iq}}) = \max_{r_{ij}} \sum_{q=1}^{m} \exp(-ED_{r_{ij} r_{iq}} * ED_{r_{iq} r_{ij}}) \tag{8}$$

其中 m 表示该聚类簇所包含的数据样本的个数；r_{jq} 表示属于该聚类簇的第 q 个数据样本在原始数据样本中的编号。

式(8)的目标是在相似矩阵 A 中寻找最优对($x_{r_{ij}}, x_{r_{iq}}$)，进而使目标函数 $f(x_{r_{ij}}, x_{r_{iq}})$取得最小值。

定义 3 在 EK-medoids 算法中，第 i 个聚类簇的新聚类中心为 $c_i = x_{r_{ij}}$，其中 $i = 1, 2, \ldots, k$，$0 \leq j \leq m$。

下面给出 EK-medoids 算法的具体步骤如下：

算法 1

输入：样本集 $X = [x_1, x_2, \ldots x_n]^{\mathrm{T}}$，初始化 K-medoids 初聚类的个数为 K。

输出：K 个样本类。

a)运用式(6)构建原始数据集的有效距离矩阵 $ED \in R^{n \times n}$，并根据有效距离矩阵构建相似矩阵 $A \in R^{n \times n}$；

b)随机选取聚类中心 $c_1, c_2, \ldots, c_k$；

c)按照式(8)逐个计算每个数据样本 x_i 与各个聚类中心的有效距离，并将该数据样本划分到与它距离最近的聚类簇中；

d)重新计算每个聚类簇的聚类中心；

e)循环 c）d)，直到聚类簇的中心点不再变化或迭代次数超过 100 次；

f)返回 $c_1, c_2, \ldots, c_k$。

2.3 基于邻域距离的属性重要性

设定一个信息系统 $S = (U, R)$，非空有限集合 U 表示对象，称为论域；非空有限集合 R 表示属性；对任意 $r \in R$ 有 $r: U \rightarrow V_r$，V_r 为属性 r 的值域；对任意 $r \in R$，$x \in U$ 有 $f(x, r) \in Vr$，$f(x, r)$是一个信息函数（或者 $f_r(x)$），该函数对 U 中对象的每个属性赋予信息值。

给定一个信息系统 $S = (U, R)$和 $x \in U$，属性集 $P \subseteq R$ 上的 x 邻域可以表示为

$$N_P(x) = \{y \mid P(x) = P(y),\ y \in U\}, \tag{9}$$

其中 $X \subseteq R$，在属性集 $P \subseteq R$ 上的 X 邻域表示为：

$$N_P(X) = \{y \mid \forall x \in X,\ P(x) = P(y),\ y \in U\}。 \tag{10}$$

定义 4 给定一个邻域系统 $S = (U, R)$，经聚类后将 U 划分成 L 个类：$C_1, C_2, \ldots, C_L$，对于任意 $P \subseteq R$，定义类信息 C 关于 P 的邻域表示为：

$$N_P(C) = \{N_P(C_1),\ N_P(C_2),\ \ldots,\ N_P(C_L)\}。 \tag{11}$$

定义 5 给定一个邻域系统 $S = (U, R)$，对象集 $X \subseteq U$ 和 Y

$\subseteq U$，在属性集 $P\subseteq R$ 上 X 和 Y 的邻域距离为

$$D_P(X,Y)=\frac{|(N_P(X)\bigcup N_P(Y))|}{|(N_P(X)\bigcap N_P(Y))|}-1\text{。} \tag{12}$$

性质 1 给定一个邻域系统 $S=(U, R)$，在属性集 $P\subseteq R$ 上 X 和 Y 的邻域距离具有单调性，对象集距离越大，其值就越大。

证明 给定一个邻域系统 $S=(U, R)$，对象集 $X\subseteq U$ 和 $Y\subseteq U$，若 $Y_1\subseteq Y_2\subseteq\ldots\subseteq U$，则有 $0\leq D_P(X, Y_1)\leq D_P(X, Y_2)\leq\ldots\leq D_P(X, U)\leq 1$。因此，可知邻域距离具有单调性，对象集距离越大，其值就越大。

定义 6 给定一个邻域系统 $S=(U, R)$，将其聚类后会得到 L 个分类：$C_1, C_2, \ldots, C_L$，进而可得属性 $r\in R$ 的属性重要度计算公式为

$$sig(r)=\frac{\sum_{i=1,\, j>i}^{L} D_r(X_j, X_i)}{C_L}\text{。} \tag{13}$$

2.4 基于 EK-medoid 聚类和邻域距离的特征选择算法

根据定义 4 和 5 可知，可以用 EK-medoids 聚类所划分类的邻域距离来计算属性重要度。由此，本文提出一种基于有效距离的聚类特征选择方法。该算法先对原始数据集运用改进的蚁群算法进行聚类，从而获得原始数据集的分类标签，在拥有分类标签的簇间用定义 6 来度量特征或属性的属性重要度，然后选择出具有较大区分度的特征子集。

如果计算高维数据的所有特征子集的分类精度，需对 2^{m-1} 个特征子集（其中 m 为数据集的属性个数）逐个进行检测，这样会使算法的时间复杂度大幅增加。于是，本文采用启发式搜索的思想来设计特征选择算法，该算法的主要思想是：以空集为出发点，每次都选择当前特征子集中属性重要度最大的属性，直到特征子集的属性重要度不会改变时终止。该算法可以使重要的属性首先加入特征子集，不会忽略重要的特性，因而，本算法选出的特征子集作为一个整体能够保持原始数据的分类能力，有效地剔除了无关的冗余属性，在此基础上设计一种基于 EK-medoids 聚类和邻域距离的特征选择算法。下面给出该算法的详细步骤。

算法 2

输入：给定一个邻域系统 $S=(U, R)$，聚类个数 k。

输出：最优或次优的特征子集。

a)在数据集上随机选择 k 个初始中心点，并使用 *EK-medoids* 聚类算法进行聚类，返回类集 C；

b)对任意 $r\in R$，计算其属性重要度 $sig(r)$；

c)$FS=\varnothing$；

d)对任意 $r_i\in R-F$，计算 $sig(FS\bigcup\{r_i\})$；

e)选择满足 $max(sig(FS\bigcup\{r_i\}))$ 的属性 r_i；

f)如果 $max(sig(FS\bigcup\{r_i\}))>0$，则

$FS\bigcup\{r_i\}\rightarrow FS$，则转向 d);否则输出 FS；

g)结束。

下面给出算法 2 的时间复杂度分析过程：

步骤 a)对原始数据集聚类的时间复杂度为 $O(knmt)$，步骤 b)计算所有特征的属性重要度的时间复杂度为 $O(km\,C_k^2)$，步骤 d)的时间复杂度为 $O(km^2\,C_k^2)$，步骤 f)的时间复杂度为 $O(km^2\,C_k^2\,m!)$，所以算法 2 总的最坏时间复杂度为 $O(knmt+O(km\,C_k^2)+O(km^2\,C_k^2)+O(km^2\,C_k^2\,m!))$。

3 实验结果与分析

3.1 实验数据与方法

为了验证算法的有效性，从 UCI 数据集中选择 3 种数据集，分别为 Chess、Lung-Cancer 和 Soybean 数据集，这 3 组 UCI 数据集的具体描述如表 1 所示。将特征选择后的数据集的聚类分析与原始数据集的聚类分析作比较来测试和验证本文算法的有效性。

表 1 三种 UCI 数据集的描述

Data Set	Samples	Attributes	Class
Chess	3196	36	2
Lung-Cancer	32	56	3
Soybean	47	35	4

为检验本论文算法的可行性，利用文献[10]提出的正确率（Accuracy, AC）、类精度（Precision, PE）和召回率（Recall, RE）等三个技术指标来对原始数据集和经过实验后的特征子集进行聚类分析，其中正确率 AC 的主要作用是验证算法的准确性，其值越大说明在聚类的过程中正确被聚类的概率越大，而类精度 PE 和召回率 RE 的主要作用是给出算法正确聚类的效率，其值越大，说明该算法聚类的效率越高。AC、PE 和 RE 的计算公式分别表示如下：

$$AC=\frac{\sum_{i=1}^{k}a_i}{n}, \tag{14}$$

$$PE=\frac{\sum_{i=1}^{k}\frac{a_i}{a_i+b_i}}{k}, \tag{15}$$

$$RE=\frac{\sum_{i=1}^{k}\frac{a_i}{a_i+c_i}}{k}\text{。} \tag{16}$$

在式(14)~(16)中，对象的总个数为 n，划分到第 i 个类的对象个数中正确的对象个数为 a_i，错误的对象个数为 b_i，而 c_i 表示应该被正确被分到第 i 个类却没有分到的对象个数，k 为聚类的个数。

3.2 实验结果分析

首先采用 EK-medoids 聚类算法对每个数据集进行聚类从而获取初始的类标签，并记录分类精度等指标；再用本文所提的特征选择算法，选择出一个特征子集，并对所选的特征子集再次聚类分析从而获取新的分类精度，最后把得到的分类精度

与第一次聚类得到的分类精度作比较。

通过对原始数据集聚类 100 次后，用聚类的 AC、PE、RE 及标准差四个指标来验证本文所提算法的有效性，其测试和实验结果如表 2~4 所示，其中，Before FS 为没有经过特征选择的聚类结果，After FS 为经过本文所提出的特征选择算法进行特征选择后，在所得到的特征子集上进行聚类分析的结果，Mean、Min、Max 和 SD 分别为各项评价指标的平均值、最小值、最大值和标准差。本文所提出的基于有效距离的 K-modes 聚类算法（简记为"算法 1"）与文献[10]采用的 K-modes 算法对三种 UCI 基因数据集（Chess、Lung-Cancer 和 Soybean）进行特征选择之前的性能比较，其实验结果如表 2~4 所示。

表 2　两种算法对数据集 Chess 在特征选择之前的实验结果

	K-modes			算法 1		
	AC	PE	RE	AC	PE	RE
Mean	0.55	0.55	0.68	0.58	0.57	0.69
Min	0.52	0.26	0.51	0.53	0.24	0.52
Max	0.71	0.71	1	0.73	0.72	0.78
SD	0.03	0.05	0.09	0.02	0.03	0.07

由表 2 的实验结果分析可知，在 Chess 数据集下，EK-medoids 聚类算法的正确率、类精度、召回率和标准差明显优于 K-medoids 算法。EK-medoids 聚类算法在召回率的标准差上略低于 K-medoids 算法。其主要原因是由于本论文的实验数据是在进行了 100 次聚类的条件下计算出的平均值，同时由于重新聚类时未被聚类的数据过多而导致的，但是 EK-medoids 聚类算法的聚类正确率和精度较高。

表 3　两种算法对数据集 Lung-Cancer 在特征选择之前的实验结果

	K-modes			算法 1		
	AC	PE	RE	AC	PE	RE
Mean	0.78	0.72	0.69	0.69	0.73	0.69
Min	0.71	0.61	0.56	0.73	0.70	0.57
Max	0.81	0.90	0.92	0.82	0.87	0.93
SD	0.01	0.03	0.07	0.02	0.03	0.04

由表 3 的实验结果分析可知，在 Lung-Cancer 数据集下，EK-medoids 聚类算法的正确率的最大值略低于 K-medoids 算法，其它指标均优于 K-medoids 算法。其主要原因是由于 Lung-Cancer 数据集本身的数据结构的稀疏性而导致的。

表 4　两种算法对数据集 Soybean 在特征选择之前的实验结果

	K-modes			算法 1		
	AC	PE	RE	AC	PE	RE
Mean	0.87	0.90	0.91	0.87	0.91	0.90
Min	0.68	0.73	0.71	0.69	0.73	0.71
Max	0.98	0.98	0.98	0.92	0.95	0.98
SD	0.10	0.07	0.07	0.10	0.08	0.07

由表 4 的实验结果分析可知，在 Soybean 数据集下，EK-medoids 聚类算法的正确率、类精度、召回率和标准差明显优于 K-medoids 算法。这充分说明了本文提出的 EK-medoids 聚类算法聚类的精度和效率明显优于文献[10]的 K-modes 算法。

由表 2~4 聚类的实验结果说明 EK-medoids 聚类算法明显优于 K-medoids 算法，从而有效地验证了 EK-medoids 聚类算法的有效性。

接下来，在表 2~4 实验结果的基础上，为了进一步验证本文所提的聚类优化特征选择算法的有效性，下面与文献[10]所提出的聚类特征选择算法比较。该算法首先在相同的三个数据集上进行 k-modes 聚类，然后经过其所提出的聚类特征选择算法进行特征选择，然后重新聚类，从而获得分类精度的指标。采用本文所提出的基于 EK-medoids 聚类和邻域距离的特征选择算法（简记为"算法 2"）与文献[10]所提出得一种基于邻域距离的特征选择算法（简记为"算法 1"），对三种 UCI 基因数据集（Chess、Lung-Cancer 和 Soybean）进行特征选择的性能比较，其实验结果如表 5~7 所示。

表 5　两种算法对数据集 Chess 特征选择的实验结果

	文献[10]的算法 1			算法 2		
	AC	PE	RE	AC	PE	RE
Mean	0.56	0.77	0.94	0.57	0.76	0.95
Min	0.52	0.76	0.82	0.53	0.76	0.83
Max	0.66	0.79	0.99	0.72	0.79	0.97
SD	0.06	0.01	0.07	0.06	0.01	0.05

表 6　两种算法对数据集 Lung-Cancer 特征选择的实验结果

	文献[10]的算法 1			算法 2		
	AC	PE	RE	AC	PE	RE
Mean	0.74	0.75	0.75	0.74	0.76	0.78
Min	0.71	0.63	0.59	0.73	0.72	0.69
Max	0.84	0.89	0.92	0.85	0.89	0.93
SD	0.04	0.05	0.09	0.03	0.06	0.08

表 7　两种算法对数据集 Soybean 特征选择的实验结果

	文献[10]的算法 1			算法 2		
	AC	PE	RE	AC	PE	RE
Mean	0.85	0.90	0.93	0.89	0.90	0.94
Min	0.72	0.63	0.59	0.73	0.73	0.72
Max	0.84	0.90	0.98	0.96	0.97	0.96
SD	0.11	0.09	0.05	0.013	0.06	0.05

由表 5~7 的实验结果分析可知，在经过算法 2 的特征选择

后重新聚类的各项指标与文献[10]所提出的算法 1 进行特征选择后重新聚类的各项指标对比，算法 2 在表 5 和表 6 中召回率的标准差略低于文献[10]所提出的算法 1，在表 7 中的召回率类精度的最大值和表 7 中类精度的标准差略低于文献[10] 所提出的算法 1。其主要原因是由于所选特征子集的结构特性而导致的。算法 2 的其它指标的实验结果均明显优于文献[10]所提出的算法 1。表 5~7 的实验结果说明了两个特征选择算法再重新聚类时，本文所提的算法 2 的各项指标均优于文献[10] 所提出的算法 1，从而反映出算法 2 所选的特征子集在分类精度和正确率方面均高于文献[10] 所提出的算法 1。这些实验结果充分表明本文所提的基于 EK-medoids 聚类和邻域距离的特征选择算法，不仅可以有效提高聚类算法的精度，还可以选择出分类精度较高的特征子集，从而验证了本文所提出的聚类优化特征选择算法的有效性和适用性。

4 结束语

本文首先用稀疏重构的方法计算数据样本之间的有效距离，将有效距离应用到聚类算法中。根据对原始数据集进行聚类划分结果所构成簇的邻域距离来确定数据集中属性的重要度计算方法，并在此基础上用启发式搜索方法设计特征选择算法；最后，在选择到的特征子集上重新聚类来验证所选特征子集的分类精度。实验结果表明，本文所提的基于 EK-medoids 聚类和邻域距离的特征选择算法，一方面提高了分类精度，另一方面也降低了计算耗时，与同类算法相比，本文所提出的算法不仅可以有效提高聚类算法的精度，还可以选择出分类精度较高的特征子集，为聚类特征选择提供了新的方法和视角。

参考文献：

[1] 徐久成，冯森，穆辉宇．基于信噪比与随机森林的肿瘤特征基因选择 [J]. 河南师范大学学报：自然科学版, 2017, 45 (2): 87-92. (Xu Jiucheng, Feng Sen, Mu Huiyu. Tumor feature gene selection based on SNR and random forest [J]. Journal of Henan Normal University: Natural Science Edition, 2017, 45 (2): 87-92.)

[2] Lai C M, Yeh W C, Chang C Y. Gene selection using information gain and improved sampled swarm optimization [J]. Neurocomputing, 2016, 16 (5): 331-338.

[3] Sun Lin, Xu Jiucheng, Tian Yun. Feature selection using rough entropy-based uncertainty measures in incomplete decision systems [J]. Knowledge-Based Systems, 2012, 36: 206-216.

[4] Wang Changzhong, Shao Mingwen, He Qiang, *et al.* Feature subset selection based on fuzzy neighborhood rough sets [J]. Knowledge-Based Systems, 2016, 111: 173-179.

[5] Chen Hongmei, Li Tianrui, Cai Yong, *et al.* Parallel attribute reduction in dominance-based neighborhood rough set [J]. Information Sciences, 2016, 373: 351-368.

[6] Liang Jiye, Wang Feng, Dang Chuangyin, *et al.* An efficient rough feature selection algorithm with a multi-granulation view [J]. International Journal of Approximate Reasoning, 2012, 53 (3): 912-926.

[7] 徐久成，黄方舟，穆辉宇，等．基于 PCA 和信息增益的肿瘤特征基因选择方法 [J]. 河南师范大学学报：自然科学版, 2018, 46 (2): 104-110. (Xu Jiucheng, Huang Fangzhou, Mu Huiyu, Wang Yun, Xu Zhanwei. Tumor feature gene selection method based on PCA and information gain [J]. Journal of Henan Normal University: Natural Science Edition, 2018, 46 (2): 104-110.)

[8] 孙林，潘俊方，张霄雨，等．一种基于邻域粗糙集的多标记专属特征选择方法 [J]. 计算机科学, 2018, 45 (1): 173-178. (Sun Lin, Pan Junfang, Zhang Xiaoyu, *et al.* A multi-label-specific feature selection method based on neighborhood rough set [J]. Computer Science, 2018, 45 (1): 173-178.)

[9] Ehsan A, Shadi M. Efficient Protocol for data clustering by fuzzy cuckoo optimization algorithm [J]. Applied Soft Computing, 2016, 41 (4): 15-21.

[10] 秦奇伟，梁吉业，钱宇华．一种基于邻域距离的聚类特征选择方法 [J]. 计算机科学, 2012, 39 (1): 175-177. (Qin Qiwei, Liang Jiye, Qian Yuhua, Clustering feature selection method based on neighborhood distance [J]. Computer Science, 2012, 39 (1): 175-177.)

[11] 孙林，刘弱南，张霄雨，等．一种基于粗糙均方残基的模糊双聚类方法 [J]. 河南师范大学学报：自然科学版, 2017, 45 (5): 93-100. (Sun Lin, Liu Ruonan, Zhang Xiaoyu, *et al.* A fuzzy biclustering approach based on rough mean square residue [J]. Journal of Henan Normal University: Natural Science Edition, 2017, 45 (5): 93-100.)

[12] 王宏杰，师彦文．结合初始中心优化和特征加权的 K-Means 聚类算法 [J]. 计算机科学, 2017, 44 (Z11): 457-459, 502. (Wang Hongjie, Shi Yanwen. K-means clustering algorithm based on initial center optimization and feature weighted [J]. Computer Science, 2017, 44 (Z11): 457-459, 502.)

[13] 马兆兴，李洪美，陈昊．应用模糊聚类的电力系统网络重构分析 [J]. 电力系统保护与控制, 2017, 45 (16): 85-89. (Ma Zhaoxing, Li Hongmei, Chen Hao. Research on network reconfiguration of power system with fuzzy clustering [J]. Power System Protection and Control, 2017, 45 (16): 85-89.)

[14] 陈龙，蔡勇，张建生，等．基于多判别参数混合方法的散乱点云特征提取 [J]. 计算机应用研究, 2017, 34 (9): 2867-2870. (Chen Long, Cai Yong, Zhang Jiansheng, *et al.* Feature point extraction of scattered point cloud based onmultiple parameters hybridization method [J]. Application Research of Computers, 2017, 34 (9): 2867-2870.)

[15] 孙义豪，李秋燕，丁岩，等．基于主成分分析及系统聚类的县域电网综合评价方法 [J]. 电力系统保护与控制, 2017, 45 (8): 30-36. (Sun Yihao, Li Qiuyan, Ding Yan, *et al.* County power grid evaluation system based on principal component analysis and hierarchical cluster analysis [J]. Power System Protection and Control, 2017, 45 (8): 30-36.)

[16] 孙胜，王元珍．基于核的自适应聚类及其在入侵检测中的应用 [J]. 计算机科学, 2008, 35 (12): 190-191. (Sun Sheng, Wang Yuanzhen. Kernel-based adaptive K-medoid clustering and its application in intrusion detection [J]. Computer Science, 2008, 35 (12): 190-191.)

[17] 何培颖，房鑫炎．基于聚类算法的关键输电断面快速搜索 [J]. 电力系

统保护与控制, 2017, 45 (7): 97-101. (He Peiying, Fang Xinyan. Fast search of the key transmission sections based on clustering algorithms [J]. Power System Protection and Control, 2017, 45 (7): 97-101.)

[18] 韩陈寿，夏士雄，张磊，等. 基于速度约束的分段轨迹聚类算法 [J]. 计算机工程, 2011, 37 (7): 219-221, 236. (Han Chenshou, Xia Shixiong, Zhang Lei, *et al.* Sub-trajectory clustering algorithm based on speed restriction [J]. Computer Engineering, 2011, 37 (7): 219-221, 236.)

[19] 张琳，陈燕，汲业，等. 一种基于密度的 K-means 算法研究 [J]. 计算机应用研究, 2011, 28 (11): 4071-4073. (Zhang Lin, Chen Yan, Ji Ye, *et al.* Research on K-means algorithm based on density [J]. Application Research of Computers, 2011, 28 (11): 4071-4073.)

[20] 赵恩来，郝文宁，赵飞，等. 改进的基于密度的航迹聚类算法 [J]. 计算机工程, 2011, 37 (9): 270-272. (Zhao Enlai, Hao Wenning, Zhao Fei, *et al.* Improved track clustering algorithm based on density [J]. Computer Engineering, 2011, 37 (9): 270-272.)

[21] Sun Lin, Zhang Xiaoyu, Xu Jiucheng, *et al.* A gene selection approach based on the fisher linear discriminant and the neighborhood rough set [J]. Bioengineered, 2018, 9 (1): 144-151.

[22] 李智远，杨习贝，徐苏平，等. 邻域决策一致性的属性约简方法研究 [J]. 河南师范大学学报：自然科学版, 2017, 45 (5): 68-73. (Li Zhiyuan, Yang Xibei, Xu Suping, Chen Xiangjian, Wang Pingxin. Attribute reduction approach to neighborhood decision agreement [J]. Journal of Henan Normal University: Natural Science Edition, 2017, 45 (5): 68-73.)

[23] Dash M, Liu J Y. Dimensionality reduction of unsupervised data [C]// Proc of the 9th IEEE International Conference on Tools with Artificial Intelligence. 1997: 532-539.

[24] 胡清华，于达仁，谢宗霞. 基于邻域粒化和粗糙逼近的数值属性约简 [J]. 软件学报, 2008, 19 (3): 640-649. (Hu Qinghua, Yu Daren, Xie Zongxia. Numerical attribute reduction based on neighborhood granulation and rough approximation [J]. Journal of Softwore, 2008, 19 (3): 640-649.)

[25] 段洁，胡清华，张灵均，等. 基于邻域粗糙集的多标记分类特征选择算法 [J]. 计算机研究与发展, 2015, 10 (1): 55-65. (Duan Jie, Hu Qinghua, Zhang Lingjun, *et al.* Feature selection for Multi-Label classification based on neighborhood [J]. Journal of Computer Research and Development, 2015, 10 (1): 55-65.)

[26] Sun Lin, Xu Jiucheng, Wang Wei, *et al.* Locally linear embedding and neighborhood rough set-based gene selection for gene expression data classification [J]. Genetics and Molecular Research, 2016, 15 (3): gmr. 15038990.

[27] Pawlak Z. Rough sets [J]. International Journal of Computer and Information Science, 1982, 11 (5): 341-356.

[28] 薛占熬，王楠，司小朦，等. 多粒度粗糙直觉模糊截集的研究 [J]. 河南师范大学学报：自然科学版, 2016, 44 (5): 131-139. (Xue Zhanao, Wang Nan, Si Xiaomeng, *et al.* Research on multi-granularity rough intuitionistic fuzzy cut set [J]. Journal of Henan Normal University: Natural Science Edition, 2016, 44 (5): 131-139.)

[29] 王思华，杨桐，段启凡，等. 基于 DT 法和粗糙集理论的接地网安全性状态评定 [J]. 电力系统保护与控制, 2017, 45 (2): 48-54. (Wang Sihua, Yang Tong, Duan Qifan, *et al.* Evaluation of security state in grounding grid based on DT method and rough set [J]. Power System Protection and Control, 2017, 45 (2): 48-54.)

[30] Brockmann D, Helbing D. The hidden geometry of complex, network-driven contagion phenomena [J]. Science, 2013, 342 (6164): 1337-1342.

[31] Xie Yuan, Zhang Wensheng, Qu Yanyun, *et al.* Discriminative subspace learning with sparse representation view-based model for robust visual tracking [J]. Pattern Recognition, 2014, 47 (3): 1383-1394.

[32] Park H S, Jun C H. A simple and fast algorithm for K-medoids clustering [J]. Expert Systems with Applications, 2009, 36 (2): 3336-3341.